

Trabajo de Fin de Grado

Integración de sistemas para el análisis de redes sociales utilizando procesamiento de lenguaje natural.

Constanza Polette León Baritussio

La Laguna, 11 de Junio de 2018

D. Dagoberto Castellanos Nieves, con N.I.F. 79234766-L profesor Doctor de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor.

C E R T I F I C A (N)

Que la presente memoria titulada:

“Integración de sistemas para el análisis de redes sociales utilizando procesamiento de lenguaje natural.”

Ha sido realizada bajo su dirección por Dña. Constanza Polette León Baritussio, con N.I.F. X-6337312-F.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos ﬁrman la presente en La Laguna a 11 de Junio de 2018.

Agradecimientos

A mis padres Sandra y Sandro por el apoyo y por siempre estar presente en los momentos difíciles. Siempre habéis estado ahí para mí.

A mi director de proyecto Dagoberto, por darme la orientación y motivación, así como las herramientas necesarias para llevar a cabo este proyecto, además de su valiosa ayuda.

A todos los profesores que embarcaron conmigo estos años universitarios, que gracias a los conocimientos brindados, he podido desarrollar tanto este trabajo como mi carrera profesional.

Finalmente a mis amigos y compañeros, por haber compartido muchos momentos. No solo habéis estado ahí para apoyarnos mutuamente y superar juntos los obstáculos, sino también por todos esos momentos de alegrías.

Licencia

\* Si quiere permitir que se compartan las adaptaciones de tu obra y NO quieres permitir usos comerciales de tu obra indica:

C:\Users\Usuario\Desktop\memoriaTFG_1415_LaTeX\tfg\images\by-nc_88x31.png

© Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial 4.0 Internacional.

**Resumen**

*El objetivo de este trabajo ha sido, desarrollar una integración de sistemas, en concreto un sistema de alerta y un sistema de análisis de sentimiento, orientado al ámbito del turismo.*

*Actualmente, las redes sociales son un medio en donde se puede acceder a la información, compartir opiniones además de crear contenido. Además, están compuestas por miles de usuarios que están en constante interacción unos con otros, por lo que se generan millones de datos. Estos datos generan una gran cantidad de información, en donde se tratan temas de actualidad, como pueden ser los destinos turísticos. Uno de los destinos más buscados y mencionados en las redes sociales por los usuarios con las Islas Canarias, que gracias a su localización geográfica, ofrecen un microclima muy agradable para los turistas. Por otro lado, además de contar con un agradable clima, Canarias posee un alto valor cultural y natural gracias sus Parques Nacionales y la sostenibilidad con el medio ambiente.*

*Para abarcar el procesamiento de todos los datos generados en las redes sociales por los usuarios, nace la necesidad de utilizar mecanismos como los sistemas de alerta y sistemas de análisis de sentimientos. Los sistemas de alerta, son sistemas que ante uno o varios eventos, son capaces de generar una notificación o alerta, para indicar a un usuario o grupo de usuarios sobre un acontecimiento. Los sistemas de análisis de sentimientos o también conocidos como “minería de opinión”, utilizan el procesamiento del lenguaje natural y la similitud semántica, para proporcionar información del estado de ánimo o grado de satisfacción de un determinado usuario. Estos sistemas, aplicados en el sector turístico, proporcionan un medio rápido y efectivo a la hora de comunicar a las personas, por ejemplo, acerca de la situación climatológica generando una alerta por oleaje en el caso de los sistemas de alerta. Y en el caso de los sistemas de análisis de sentimientos, opiniones positivas o negativas dependiendo del léxico empleado, basándose en las propias experiencias de los usuarios.*

**Palabras clave:** Red Social, sistema de alerta, sistema de análisis de sentimiento, turismo, integración.

**Abstract**

*The objective of this work has been,the extraction of data from the social network Twitter,it is a perfoming tool capable of analytics on the semantic similaritis between the concepts used on social networks.*

***Keywords:*** *Social network, semantic similarity.*

Índice General

[Capítulo 1 9](#_Toc516482336)

[Introducción 9](#_Toc516482337)

[1.1 Internet 10](#_Toc516482338)

[1.2 Web 2.0 11](#_Toc516482339)

[1.3 Redes Sociales 12](#_Toc516482340)

[1.4 Redes Colaborativas 17](#_Toc516482341)

[1.4.1 Introducción y definición 17](#_Toc516482342)

[1.4.2 Tipologías de redes organizacionales 18](#_Toc516482343)

[1.4.3 Similitudes y diferencias con respecto a las redes sociales 19](#_Toc516482344)

[1.5 Tecnologías 20](#_Toc516482345)

[1.5.1 Procesamiento del lenguaje natural (PLN) 20](#_Toc516482346)

[1.5.2 RStudio. 22](#_Toc516482347)

[1.5.3 Lenguaje R 23](#_Toc516482348)

[1.5.4 ShinyApps. 25](#_Toc516482349)

[1.5.5 GitHub 25](#_Toc516482350)

[1.6 Objetivos del proyecto 26](#_Toc516482351)

[Capítulo 2 27](#_Toc516482352)

[Sistemas de alerta y sistemas de análisis de sentimientos. 27](#_Toc516482353)

[2.1 Sistemas de alerta. 27](#_Toc516482354)

[2.1.1 Sistemas de alertas de vigilancia tecnológica 27](#_Toc516482355)

[2.1.2 Sistemas de Alerta Temprana 28](#_Toc516482356)

[2.2 Sistemas de análisis de sentimientos 29](#_Toc516482357)

[2.3 Propuesta de proyecto 30](#_Toc516482358)

[2.3.1 Red social del proyecto. 30](#_Toc516482359)

[2.3.2 Nuestra Propuesta 30](#_Toc516482360)

[Capítulo 3 32](#_Toc516482361)

[Análisis de requisitos y caso de uso. 32](#_Toc516482362)

[3.1 Requisitos 32](#_Toc516482363)

[3.1.1 Servidor 32](#_Toc516482364)

[3.1.2 Cliente 33](#_Toc516482365)

[3.2 Diagramas de casos de uso 34](#_Toc516482366)

[Capítulo 4 35](#_Toc516482367)

[Integración de los sistemas de alerta y de análisis de sentimientos propuestos. 35](#_Toc516482368)

[4.1 Selección de datos. 35](#_Toc516482369)

[4.2 Extracción de datos. 35](#_Toc516482370)

[4.3 Tratamiento de los datos obtenidos 36](#_Toc516482371)

[4.4 Análisis de los datos 37](#_Toc516482372)

[4.5 Alertas enviadas 38](#_Toc516482373)

[4.6 Librerías y entorno de desarrollo 38](#_Toc516482374)

[5.1 La aplicación con shinyapp 39](#_Toc516482375)

[Capítulo 5 40](#_Toc516482376)

[Conclusiones y líneas futuras. 40](#_Toc516482377)

[5.1 Conclusiones 40](#_Toc516482378)

[5.2 Líneas Futuras 42](#_Toc516482379)

[Capítulo 6 42](#_Toc516482380)

[Capítulo 7 43](#_Toc516482381)

[Presupuesto 43](#_Toc516482382)

[7.1 Presupuesto del trabajo realizado 43](#_Toc516482383)

[Referencias 46](#_Toc516482384)

[Apéndice A 52](#_Toc516482385)

[Conexión a Twitter con OAuth 52](#_Toc516482386)

[Apéndice B 52](#_Toc516482387)

[Extracción de datos 52](#_Toc516482388)

Índice de figuras

[Figura 1. Comparativa Deportes. 16](#_Toc516565373)

[Figura 2. Comparativa Universidades y facultades 17](#_Toc516565374)

[Figura 3. Comparativa Gobierno 18](#_Toc516565375)

[Figura 4: Caso de uso del servidor. 35](#_Toc516565376)

[Figura 5: Caso de uso del cliente 35](#_Toc516565377)

[Figura 6: Código para autenticación con OAuth. 37](#_Toc516565378)

[Figura 7: Dataframe de extracción de datos 38](#_Toc516565379)

[Figura 8: Datos en fichero CSV 39](#_Toc516565380)

[Figura 9: Limpieza de datos. 39](#_Toc516565381)

[Figura10: Código para envío de alertas en el servidor. 40](#_Toc516565382)

[Figura 11: Código para visualizar el formulario de la alerta. 41](#_Toc516565383)

Índice de tablas

[Tabla 7.1. Tabla resumen de los Tipos. 44](#_Toc516481864)

## Capítulo 1

## Introducción

Internet, es una herramienta imprescindible en la sociedad actual, ya que, crea puentes de comunicación entre los individuos, eliminando así las barreras de espacio y tiempo. Con la llegada de la Web 2.0 (término acuñado por Tim O`Reilly ), que está basada en comunidades virtuales de usuarios, en donde estos pueden colaborar e interactuar entre sí, además de crear contenidos. Y una gama especial de servicios, como son las redes sociales los blogs y las wikis.

La creación de las redes sociales supone una nueva forma de comunicación. Las personas pueden expresar opiniones digitales en foros, debates y hacer que estas lleguen a un público muy amplio y diverso.

Dado que en las redes sociales hay un contínuo intercambio de información, los métodos tradicionales empleados para el tratamiento de los datos, ya no resultan efectivos a la hora de procesar estos datos. No son capaces de afrontar los grandes volúmenes de datos. Existen diversas investigaciones realizadas en el ámbito de las redes sociales para intentar solventar este problema. Estos trabajos, hacen uso de métodos más actuales para el tratamiento de los datos, por ejemplo, mediante el procesamiento del lenguaje natural (PLN), son capaces de realizar nuevos métodos para el análisis de los datos obtenidos de las redes sociales.

La memoria está compuesta por los siguientes cuatro capítulos: 1. Introducción, en este capítulo se recoge, los conceptos clave, como: Internet, Web 2.0, redes sociales, redes colaborativas. Tecnologías necesarias para llevar a cabo el proyecto: Procesamiento del lenguaje natural, lenguaje de programación R y sus librerías, entornos de desarrollo como Rstudio y Shinyapp. Así como, una descripción de la propuesta para el Trabajo de Fin de Grado (TFG) y los objetivos que se quieren alcanzar. 2. Sistemas de alerta y sistemas de análisis de sentimientos, pretende definir en qué consiste cada uno de estos sistemas y exponer ejemplos sobre en qué ámbitos se están desarrollando y su utilidad. 3. Integración de sistemas de alerta y sistemas de análisis de sentimientos, la intención de este capítulo es explicar cómo y con qué metodologías, se realizó la elaboración de una herramienta, que permite la integración de un sistema de alerta y un sistema de análisis de sentimientos. 4. Caso de uso para la integración de los sistemas de alerta y de sentimientos. En este capítulo se detallara un ejemplo en donde se podría utilizar el sistema planteado. 5 y 6. Recomendaciones y conclusiones, con este capítulo se recogen las posibles mejoras que se podrían implementar para tener una herramienta más óptima, así como las conclusiones obtenidas después de la realización del proyecto. 7 Presupuesto, capítulo en el cual se coteja una estimación aproximada del coste del proyecto, en base a las horas empleadas, software utilizado y el hardware.

### Internet

El nombre Internet, proviene de las palabras en inglés “Interconnected Networks”, que significa “redes interconectadas”. Sus orígenes se remontan a 1969, cuando se estableció la primera conexión entre ordenadores, conocida como ARPANET.

Internet, es una arquitectura de sistema que ha revolucionado las comunicaciones y los métodos de comercio. Permite que varias redes de computadoras estén interconectadas. Proporciona un mecanismo potente y accesible por cualquier individuo, ya que, ofrece una amplía información que puede ser utilizada para cualquier propósito en general. La difusión de textos digitalizados, imágenes y grabaciones de audio y vídeo, ha dado lugar a una explosión de información [1].

El internet, se ha convertido en uno de los medios más importantes en la sociedad desde finales del siglo 20 y principios del 21.Trajo consigo una nueva forma de comunicación entre las personas y una mejora en la transmisión y acceso de la información eliminando así las barreras de espacio/tiempo. Surge así la World WideWeb [2] o también conocida como Web 1.0. Mediante el concepto denominado la “webvolución” [3], en otras palabras, la evolución de la web, el enfoque del concepto de la web ha ido transformándose de “acceso y hallazgos” (Web 1.0), a “compartir, participar y colaborar”(Web 2.0), hasta llegar a la “colaboración inmersiva y co-creación”(Web 3.0) [4] [5].

### Web 2.0

Después de la llegada de la World Wide Web en los 90 y la Web2.0 [6], para referirse a una segunda generación en la historia de la Web. La Web 2.0 se puede entender cómo –“todas aquellas utilidades y servicios de Internet que se sustentan en una base de datos, la cual puede ser modificada por los usuarios del servicios, ya sea en su contenido (añadiendo, modificando o borrando información o asociando datos a la información existente), pues bien en la forma de presentarlos, o en contenido y forma simultáneamente”- [7]. Se puede afirmar que el término “Web 2.0” representa la evolución de las aplicaciones tradicionales hacia aplicaciones web enfocadas en el usuario final. No se trata de una nueva tecnología sino de una actitud de colaboración y participación de las personas para proporcionar mejores datos, nuevos servicios y aplicaciones on-line [8]. En donde los usuarios pueden colaborar e interactuar entre sí [9], creando contenidos dentro de una comunidad virtual y conectándose con otros individuos.

Aunque, todos los sitios de la Web están interconectados entre sí, con el fin de organizar los recursos digitales [9], proponen una taxonomía compuesta por cuatro líneas fundamentales que ayudarán al usuario en la selección de aquellas aplicaciones que cubran de una mejor manera sus necesidades e intereses. La estructura propuesta se denomina “los cuatro pilares de la Web 2.0”, que son [10].

* **Redes sociales.** Describe todas aquellas herramientas diseñadas para la creación de espacios, que promuevan o faciliten la formación de comunidades e instancias de intercambio social.
* **Contenidos.** Hace referencia a aquellas herramientas que favorecen la escritura en línea, así como su distribución e intercambio.
* **Organización social e inteligente de la información.** Herramientas y recursos para etiquetar, sindicar e indexar, que faciliten el orden y almacenamiento de la información, así como otros recursos disponibles en la Red.
* **Aplicaciones y servicios (mashups).** Dentro de esta clasificación, se incluye numerosas herramientas software, plataformas en línea y un híbrido de recursos creados para ofrecer servicios de valor al usuario final.

### Redes Sociales

Como se describía en la sección anterior, la Web 2.0 está compuesta por comunidades de usuarios y servicios, como: las redes sociales, los blogs y las wikis.

Para definir a las redes sociales, existen diferentes prospectivas: las que se centran en la comunicación y colaboración o las que se basan en estudio de las características estructurales de sus herramientas [11]. Algunos autores como [12], definen a las redes sociales como: herramientas que facilitan el construir un perfil público dentro de un sistema delimitado, pudiendo así, articular una lista de diferentes usuarios con los que comparte una conexión, y además poder ver las conexiones que suceden dentro de su sistema. Otros autores [13], las define como: una estructura social que se puede representar como nodos conectados por aristas, en donde los nodos representan a los individuos y las aristas las relaciones existentes entre ellos. Para el autor [14], se centra en un entorno educativo, señala que son un entorno que permite la discusión de diferentes problemáticas en espacios interactivos y flexibles de aprendizaje. Nosotros definiremos a las redes sociales como una estructura compuesta por un grupo de individuos u organizaciones y que están relacionados entre sí, ya sea, mediante relaciones laborales, amistad, familiar. Además, de un punto de encuentro en donde los usuarios comparten, crean y acceden a la información.

En la sociedad actual, las redes sociales despiertan un inusitado interés social y profesional. En el ámbito académico, el interés se centra en conocer los efectos que estas provocan en el éxito y desarrollo educativo de cualquier estudiante, sobre todo a niveles universitarios [15]. En el resto de ámbitos como el laboral, social y relaciones afectivas, la creación de las redes sociales supone una nueva forma de relacionarse. Estas herramientas ofrecen un espacio virtual para que, miles de personas puedan compartir contenidos multimedia con intereses similares [16] . Se consideran puntos de encuentro, en los que es posible acceder a la información, compartir impresiones, consultar archivos y recursos disponibles a tiempo real [17].

Redes sociales como, Facebook, Twitter, Orkut, son estructuras compuestas por una infinidad de usuarios relacionados entre sí, independientemente que se conozcan o no. En estas redes sociales, al tener un gran número de usuarios, se manejan grandes volúmenes de datos, ya que, los individuos comparten, buscan y difunden información a través de texto, imágenes o audio. Gracias a las redes sociales, las personas pueden expresar opiniones digitales en foros, debates y hacer que estas lleguen a un público muy amplio y diverso.

En el año 2002, comenzaron a aparecer los primeros sitios web, promocionando las redes de círculos de amigos en línea. La popularidad de estos sitios creció rápidamente y compañías como por ejemplo, Myspace y Linkedin surgieron en el año 2003, Google lanzó Orkut en 2004 y en el mismo año se dio a conocer Facebook, en 2006 se lanzó Twitter.

Uno de los objetivos de la creación de estas redes, es crear un sitio de interacción virtual en donde miles de personas de todo el mundo puedan compartir, comunicarse y cooperar entre sí.

**Clasificación de las redes sociales.**

Las redes sociales según su tipología pueden clasificarse entre redes sociales horizontales o generales y verticales o especializadas.

**Horizontales.**

Las redes sociales horizontales, no tienen una temática definida, están orientadas al público en general y se centran en las relaciones de contacto entre usuarios. Provee herramientas para que las personas se relacionen entre sí y todas comparten características: crear un perfil, compartir contenidos, generar lista de contactos [18]. Algunas de ellas son: Facebook, Orkut, Google+.

**Verticales.**

En cuanto a las redes sociales verticales, tienden a especializarse según los intereses de sus usuarios, está orientada a un público específico y diariamente ganan miles de seguidores. Cualquier red social que se centre en un tema en concreto, puede dar origen a una nueva categoría de clasificación y es por ello que destacaremos las más relevantes, como por ejemplo las redes sociales según su temática, actividad y por ultimo según el contenido compartido [18].

**Temática.**

**Profesionales**. Son redes enfocadas en negocios y actividades comerciales. Los usuarios definen un perfil en donde detallan su ocupación, estudios académicos, experiencia laboral. La más conocida es Linkedin.

* **Viajes**. Estas redes sociales, conectan viajeros que comparten experiencias, sirviendo así de guía para otros viajeros. Un ejemplo de ellas sería Travellerspoint, tripadvisor.
* **Relaciones personales**. Son redes sociales propias para relaciones personales de pareja, por ejemplo: meetic, Tinder.

**Actividad.**

* **Microblogging**. Estas redes sociales, ofrecen un servicio de envío y publicación de mensajes breves de texto. También permiten seguir a otros usuarios, aunque esto no establece necesariamente una relación recíproca, como los seguidores. La red social más destacada es Twitter.
* **Geolocalización.** Permiten, mostrar la localización de una persona o establecimiento. Mediante estas redes sociales los usuarios pueden localizar el contenido digital que comparten. Por ejemplo: Foursquare, Trippy.

**Contenido.**

* + **Fotos.** Estos servicios, ofrecen la posibilidad de almacenar, ordenar, buscar y compartir fotografías. Las más importantes en número de usuarios son: Instagram, Flickr, Fotolog.
  + **Vídeos.** Los sitios web de almacenamiento de vídeos, se han popularizado de tal manera, que en los últimos años incorporan la creación de perfiles y listas de amigos para la participación colectiva mediante los recursos de los usuarios, y los gustos sobre los mismos. Algunos son: Youtube, Dailymotion.
  + **Documentos.** Por la red navegan documentos de todo tipo en formatos diversos, en estas redes sociales podemos encontrar, publicar y compartir los textos definidos por nuestras preferencias de una manera fácil y accesible. Su mayor exponente es Scribd, GoogleDrive.

Nosotros, haremos una clasificación por medio de gráficas para representar a las redes sociales más importantes en la red. Pondremos énfasis en lo que respecta a la compartición de información mediante la opinión de los usuarios. Y a los temas tratados con especial relevancia como por ejemplo temas de actualidad que son: la política, Internet y telecomunicaciones, destinos turísticos, noticias en general, deportes y así una infinidad de temáticas.

Según las estadísticas del último año, en todo el mundo a lo que deportes se refiere, Twitter es la más utilizada a la hora de divulgar opiniones acerca de los partidos, competiciones y campeonatos.

Deportes



Figura 1. Comparativa Deportes.

En donde usuarios comparten sus opiniones, en partidos, competiciones y campeonatos realizados en el último año.

En el caso de las facultades y universidades que utilizan redes sociales, la más destacada es Twitter frente a redes sociales como Orkut o Instagram.

Facultades

Y

Universidades



Figura 2. Comparativa Universidades y facultades

Comparativa Universidades y facultades, de los últimos cinco años en donde la red social más utilizada es Twitter frente a Orkut o Instagram.

Comparativa en España de las redes de mensajería rápida más utilizadas, a la hora de compartir contenidos a cerca del gobierno español.

[](https://trends.google.es/trends/explore?cat=19&date=today%203-m&geo=ES&q=twitter,telegram,/m/0gwzvs1)

Gobierno

Figura 3. Comparativa Gobierno

Comparativa Gobierno, en donde los el medio de comunicación de mensajería corta más utilizado es Twitter en comparación con Telegram o Whatsapp.

### Redes Colaborativas

#### Introducción y definición

En la sección anterior, se ha definido el concepto de redes sociales y sus clasificaciones. En este apartado se definirá que es una red colaborativa y su similitud, diferencias con respecto a las redes sociales.

Las Redes Colaborativas (RC) tienen una especial importancia. Podemos definir a una Red Colaborativa como una red compuesta por varias entidades autónomas, dispersas geográficamente y diferentes en términos de: entorno, cultura, capital social y objetivos. En las RC, la colaboración entre los miembros de la red hace posible alcanzar los objetivos marcados y que de manera independiente, estos objetivos serían más costos de alcanzar para las organizaciones [19]. Otros autores como [20], describe a las RC como, una red que se compone de un conjunto de vínculos, que pueden representar relaciones, o ausencia de ellas. Los nodos pueden ser actores, unidades organizativas y organizaciones. El objeto de intercambio define el tipo de red a estudiar (bienes, servicios, influencias, afecto).Las relaciones se asumen como perdurables [21].

#### Tipologías de redes organizacionales

En las redes colaborativas, identifican cuatro escenarios posibles en los que analizar los patrones de relaciones entre organizaciones, clasificados en [22]: Relación binaria, Relación estelar, Red de trabajo, entre redes. Nosotros haremos una segunda clasificación: análisis centrado en una organización y análisis orientado a la red.

1. **Análisis centrado en una organización.**
   1. **Relación binaria**. El análisis de la relación se realiza respecto a dos entidades que establecen un acuerdo de colaboración. El contenido de la relación se centra en definir:
      1. Predisposición, o no, para actuar de manera conjunta, mediante el intercambio de bienes o servicios.
      2. Dependencia mutua. Resultado de la acción conjunta de las dos organizaciones.
      3. Organización del vínculo. Representa el grado de colaboración que puede medirse en términos de: formalización del acuerdo, intensidad, reciprocidad y estandarización.
   2. **Relación estelar**. Es un modelo centralizado en el que la organización que predomina, analiza de manera dual, su actividad de cooperación y colaboración con cada uno de las otras organizaciones o entidades.
2. **Análisis Orientado a la red.**
   1. **Red de trabajo.** Se trata de un conjunto de organizaciones que se movilizan de manera cooperativa con la finalidad de alcanzar un objetivo determinado aunque para lograrlo no necesariamente deben adquirir una entidad. La duración de esta colaboración, depende del objetivo consensuado y predeterminado cuyo alcance de manera individual no sería posible. Este tipo de colaboración, no se enfoca hacia las relaciones entre los participantes, sino en el comportamiento organizativo global. Para conseguirlo, los participantes deben determinar [23].
      1. El número de actores implicados.
      2. El ámbito de poder al que accede la organización escogida como líder.
      3. La similitud de valores y actitudes entre los miembros
      4. El efecto que tiene el comportamiento de otros equipos frente a éste.
   2. **Entre Redes.** La colaboración entre redes, representa a un conjunto de redes, generalmente con un número elevado de integrantes, que mantienen algún tipo de relación entre ellas. Difiriendo con la red de trabajo, la colaboración entre redes, posee una naturaleza dinámica y su extensión temporal puede ser indefinida [24].

#### Similitudes y diferencias con respecto a las redes sociales

Anteriormente, hemos definido a las redes sociales como una estructura compuesta por un grupo de individuos u organizaciones y que están relacionados entre sí, y en donde los usuarios comparten, crean y acceden a la información. Las redes colaborativa al igual que las redes sociales, son una estructura compuesta por personas u organizaciones que colaboran entre sí ya sea en algún proyecto, investigación para lograr objetivos, comunes, independientemente del ámbito en donde se desarrollen, en donde interactúan mediante los ordenadores.

La diferencia entre las redes sociales con respecto a las redes colaborativas, es que, las redes colaborativas se crean con el propósito de unir a personas, u organizaciones, para que colaboren entre sí, mediante el proceso colaborativo compuesto por: colaboración, cooperación, coordinación y networking. Con un fin en común “alcanzar un objetivo final” que puede ser resolver un problema o algún proyecto de investigación. En cambio las redes sociales, simplemente se crean para que las personas se creen un perfil en la red y así conectar usuarios, compartir contenido si así lo quieren y difundirlo.

### Tecnologías

En las secciones anteriores, se ha definido a las redes sociales, a las redes colaborativas, y sus respectivas diferencias. Además, en cómo estás están compuestas por miles de usuarios. Las redes sociales, al tener un gran número de usuarios, están constantemente compartiendo y difundiendo información a través de imágenes, texto o audio. Por lo que se generan grandes volúmenes de datos. Para poder analizar esa gran cantidad de datos, es necesario la utilización de tecnologías como: el procesamiento del lenguaje natural (PLN), lenguajes de programación como el lenguaje R y entornos de desarrollo como Rstudio y Servicios para el despliegue de aplicaciones como Shinyapp.

#### Procesamiento del lenguaje natural (PLN)

Para entender el procesamiento del lenguaje natural, primero definiremos en que consiste el lenguaje natural (LN). Un lenguaje natural es aquel que ha evolucionado con el tiempo para fines de comunicación humana, como el español o alemán [25]. El lenguaje natural, es la vía de comunicación que utilizan las personas para comunicarse entre ellas habitualmente.

El Procesamiento de Lenguaje Natural (PLN), se entiende como la habilidad que tiene una máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje.Un traductor autómatico sería un ejemplo de dispositivo o software de PLN [26].

El PLN tiene aplicaciones muy variadas, ya que su alcance es muy extenso, algunas aplicaciones son [25].

* Traducción automática.
* Recuperación de la información.
* Extracción de Información y Resúmenes.
* Tutores inteligentes.
* Reconocimiento de Voz.

La arquitectura de un sistema de PLN, se fundamenta en una definición del LN por niveles, que son: fonológico, morfológico, sintáctico, semántico y pragmático [25].

* **Nivel Fonológico**. Trata de cómo las palabras se relacionan con los sonidos que representan.
* **Nivel Morfológico**. Trata de cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.
* **Nivel sintáctico.** Trata de cómo las palabras pueden unirse para formas oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.
* **Nivel semántico:** trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de las oraciones aisladas.
* **Nivel Pragmático.** Trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

Existen cuatro componentes necesarios para usar y manejar de manera más provechosa el conocimiento plasmado en textos. Estos son: la digitalización, la recuperación, la presentación de la información y su uso directo por el software [26].

1. **Digitalización de la información.** Existe muchísimos documentos, libros, manuales en donde su contenido está plasmado en papel y es difícil de acceder a la información. Por ello es importante que a la hora de usar el PLN el contenido este en formato digital. Por digitalización podemos entender como la obtención de la información en formato digital en donde el texto es una secuencia de letras, por ejemplo los correos electrónicos, documentos en periódicos digitales, información en las redes, etc.
2. **Recuperación de la información**. También conocida como búsqueda de la información, tiene la problemática que un mismo concepto o idea puede expresarse de maneras distintas, utilizando diferentes palabras. Por ejemplo: el usuario expresa su interés con la frase “la derrota del Madrid” y el documento relevante para tal petición es “la victoria del Barcelona”. Los dos textos no tienen ninguna palabra en común, pero un humano, usando su experiencia lingüística (derrota—victoria) y su conocimiento del mundo (Madrid—Barcelona) fácilmente detectaría la relevancia del documento para la petición. Se han alcanzado avances significativos para que los programas puedan usar este tipo de razonamiento para así, satisfacer de una mejor manera las necesidades de los usuarios.
3. **Presentación de la información**. Es importante que la información contenida en los textos, este representado de una manera eficaz. Un ejemplo sería, la construcción automática de resúmenes, es decir, un generador automático de resúmenes, a partir de un texto largo, detectar lo más transcendente y presentarlo en un texto corto que se pueda leer en un tiempo más razonable. Otra forma de resumir la información para hacerlos más manejables sería, agrupar y clasificar los documentos que sean semejantes entre sí. La minería de texto, es otra manera de resumir el contenido textual (encontrar opiniones expresadas en los textos, las tendencias de cambio de estas opiniones o las relaciones inesperadas entre los eventos descritos en los textos), la extracción de información (llenar bases de datos sobre un tema específico, leyendo los textos) y sistemas de soporte a la toma de decisiones (buscar, sintetizar y presentar de manera eficiente la información relevante para un directivo).
4. **Uso de la información**. El uso de la información contenida en los textos por el mismo software para resolver tareas más complejas, permite a la máquina encontrar el conocimiento necesario de los textos disponibles. Por ejemplo: los artículos científicos o los libros de texto.

#### RStudio.

Rstudio, es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, un editor de resaltado de sintaxis, que soporta la ejecución de código directo, así como herramientas para: la depuración y gestión de espacio de trabajo. Rstudio está disponible para: Windows, Mac y Linux se puede ejecutar en el escritorio o en un navegador conectado a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux) [27].

Rstudio tiene como objetivo proporcionar un entorno de desarrollo informático estadístico R. Permite el análisis y desarrollo para que se lleven a cabo los análisis correspondientes con el lenguaje R. Las principales características son:

* Acceso Rstudio localmente.
* El resaltado de sintaxis, completado de código, y la sangría inteligente.
* Ejecutar código R directamente desde el editor de código fuente.
* Saltar rápidamente a la definición de funciones.
* Gestionar fácilmente múltiples directorios de trabajo utilizando proyectos.
* R ayuda integrada y documentación.
* Depurador interactivo para diagnosticar y corregir los errores rápidamente.
* Paquete de herramientas de desarrollado por usuarios.

#### Lenguaje R

El lenguaje R, además de ser un lenguaje de programación, también es un entorno con un enfoque al análisis s y gráficos. Es un proyecto GNU similar al lenguaje de programación S. Fue desarrollado en Bell Laboratories anteriormente AT & T, ahora Lucent Technologies), por John Chambers y colegas. R puede considerarse como una implementación diferente a S. Entre estos dos lenguajes de programación, existen algunas diferencias importantes, aunque mucho código escrito para S se ejecuta bajo R.

R, proporciona una gran variedad de modelos estadísticos (modelos lineales y no lineales, pruebas estadísticas clásicas, clasificaciones, agrupaciones, etc.) y técnicas graficas además es muy extensible. El lenguaje R proporciona facilidad a la hora de crear publicaciones bien diseñadas incluyendo símbolos matemáticos y formulas donde se es necesario. Además es uno de los lenguajes más utilizados por la comunidad de programadores, y es muy popular en campos como: la minería de datos, investigaciones biomédicas, la bioinformática, y finanzas. R está disponible como software libre bajo la licencia de GNU, compila y se ejecuta en una amplia variedad de plataformas como: UNIX, Windows, MacOS. Ofrece la posibilidad de trabajar con distintas bibliotecas o librerías que ofrecen funciones para analizar los datos que se le provee [28].

Características.

R es un conjunto integrado de paquetes para la manipulación de datos, el cálculo y la visualización gráfica [29].Incluye:

* Una instalación eficaz de tratamiento y almacenamiento de datos.
* Una serie de operadores para cálculos matemáticos.
* Una gran colección de herramientas integradas para el análisis de datos.
* Instalaciones gráficas para el análisis de datos y visualización.
* Proporciona gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
* Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas.
* Permite que los usuarios lo extiendan definiendo sus propias funciones.
* Orientado a objetos, característica que hereda de S.

Librerías.

R, es miembro de un proyecto colaborativo y abierto, en donde sus usuarios pueden publicar paquetes y extender su configuración básica. R se puede extender fácilmente con los paquetes, existen unos ocho paquetes suministrados con la distribución de R y muchos de ellos están disponibles en CRAN (<https://www.r-project.org/>).

Librerías para el proyecto:

En nuestro proyecto para su la elaboración utilizaremos algunas librerías como:

* **library (“twitteR”):** Proporciona una interfaz a la API web de Twitter.
* **library (“tm”):** Proporciona un marco para aplicaciones de minería de texto dentro de R.
* library (“ROAuth”): Proporciona una interfaz de Oauth 1.0, permitiendo así a los usuarios de Twitter autenticarse vía Oauth.
* library (“streamR”): Este paquete ofrece una serie de funciones que permiten a los usuarios acceder por filtro a los usuarios de Twitter.
* Shiny: Shiny es un paquete R de código abierto que facilita la creación de aplicaciones web interactivas directamente desde R. Puedes alojar aplicaciones independientes en una página web o insertarlas en documentos de R Markdown o crear cuadros de mando. También puede extender sus aplicaciones Shiny con temas CSS, html widgets y acciones de JavaScript.

#### ShinyApps.

ShinyApps, es un servicio de hosting utilizado para el despliegue de aplicaciones Shiny. Existen otras posibilidades para llevar a cabo el despliegue como:

* En la máquina local mediante el IDE Rstudio y bajo cualquier sistema operativo.
* En un servidor dedicado con el Server Shiny de Rstudio.
* En algún servicio de hosting dedicado especializado para Shiny.

Para este proyecto se utilizara el servicio de hosting ofrecido por ShinyApps, ya que, es fácil de utilizar y no supone coste alguno. Solo se necesita una cuenta de usuario para hacer el despliegue. Además la integración del paquete Shiny y Rstudio con ShinyApps, permite realizar el despliegue de forma automática, sin tener que configurar y desplegar de forma manual [30].

#### GitHub

GitHub es una plataforma Web de desarrollo colaborativo y alojamiento de repositorios basada en Git (software de control de versiones). Ofrece toda la funcionalidad de Git e integra diversas herramientas de control de acceso, colaboración, trazabilidad, gestión de tareas y control de proyectos. Además, proporciona una interfaz Web que permite al usuario registrado crear repositorios vacíos o por clonación de otro repositorio hospedado en GitHub (fork en la terminología de GitHub), enviar solicitudes de cambio entre repositorios hospedados (pull request en la terminología de GitHub), y gestionar dichas solicitudes. Asimismo, los repositorios hospedados en GitHub pueden actuar como repositorios remotos de repositorios locales. [31]

### Objetivos del proyecto

Los objetivos que se persiguen con este proyecto son los siguientes:

(OB1). Estudiar del estado del arte de las redes sociales.

(OB2). Analizar distintas metodologías para sistemas de alerta y sistemas de análisis de sentimientos.

(OB3). Diseñar e implementar un sistema de alerta y un sistema de análisis sentimientos con los datos propuestos.

(OB4). Testear, validar e integrar en una sola herramienta, un sistema de alerta y un sistema de análisis de sentimientos.

(OB5). Proponer un caso de uso para la herramienta que tendrá integrada un sistema de alerta y un sistema de análisis de sentimientos.

(OB6). Elaborar documentación técnica.

## Capítulo 2

## Sistemas de alerta y sistemas de análisis de sentimientos.

En este capítulo, definiremos en qué consiste un sistema de alerta y un sistema de análisis de sentimientos, mostrando ejemplos de cada uno de ellos. Y por último detallaremos el sistema que queremos proponer en nuestro proyecto.

### 2.1 Sistemas de alerta.

Un sistema de alerta, se podría definir como: sistema que ante uno o varios eventos, es capaz de generar una notificación o alerta, para indicar a un usuario o grupo de usuarios sobre un suceso, acontecimiento que está ocurriendo en ese momento o simplemente retornar una información.

Los sistemas de alerta han sido implementados en diversos campos como: la vigilancia tecnológica, la climatología (SAT), educación, sanidad, entre otros.

#### 2.1.1 Sistemas de alertas de vigilancia tecnológica

En un sistema de alertas de vigilancia tecnológica, como por ejemplo “ALERTAS”, ofrece información de actualidad sobre: ciencia, tecnología e innovación. A través de estas alertas, los usuarios pueden tener acceso a: eventos actualizados, noticias diarias, artículos recientes, ferias y congresos. Este sistema para poder ofrecer el servicio de alerta, necesita que el usuario este registrado y que además en el formulario de registro, especifique sus preferencias. Posteriormente el sistema, recogerá está información para llevar a cabo y de manera diaria el proceso de detección, clasificación y priorización de resultados para así, ofrecer al usuario un resumen de noticias, eventos y novedades.

El estos sistemas, se incorpora tecnologías del lenguaje natural para la recuperación, extracción y clasificación de información documental, desarrolladas bajo los estándares de la web semántica. Con la suma de estrategias y técnicas es posible presentar al usuario solo aquella información que cubra sus necesidades. Estas son: la creación y mantenimiento de ontologías, la asignación semi-automática de metadatos y generación de vinculaciones y relaciones semánticas entre estos para la extracción de información textual.

Permiten, la clasificación de la información procedente de fuentes relevantes en el ámbito científico. Extrae solamente la información relevante dependiendo de las preferencias de los usuarios y la almacena en bases de datos para su posterior procesamiento. Está compuesto por varias partes [32].

* Clasificación inteligente. Consiste en hacer que el sistema aprenda a diferenciar entre la información relevante de la que lo no es. Y para ello utiliza técnicas de aprendizaje automático y categorización de textos, y a diferenciar entre unas temáticas y sectores de otros.
* Rastreo automático. El sistema tecnológico rastrea periódicamente una selección de dominios y fuentes de noticias especializadas. Cuando se detectan nuevos contenidos, el sistema los descarga, categoriza automáticamente y selecciona los resultados relevantes, descartando los no relevantes. Estos contenidos actualizados y relevantes son las denominadas Alertas de Vigilancia. Éstas son almacenadas en el escritorio de la plataforma virtual y accesible para todos los usuarios, previamente registrados. Cada usuario recibirá entonces un conjunto de alertas específicas y personalizables según sus preferencias seleccionadas.
* Ranking y envío por correo. Según el medio de consulta y la frecuencia escogida, el usuario recibirá un resumen personalizado con las diez alertas más relevantes del conjunto de todas las alertas que el sistema tecnológico ha localizado, clasificado y priorizado. Para ello, se elabora un ranking automático por usuario sector de actividad, países de interés y temáticas escogidas en su perfil.

#### 2.1.2 Sistemas de Alerta Temprana

Actualmente, se está viviendo en una época en donde se están sufriendo cambios climáticos a nivel global, en donde los distintos países del mundo están intentando adaptarse a las nuevas condiciones climatológicas.

Un Sistema de Alerta Temprana (SAT), se puede definir como un sistema de colección de información variada que, mediante monitoreo constante, permite advertir sobre situaciones amenazantes a la seguridad alimentaria y a la seguridad civil [33].

El objetivo principal de un sistema de alerta temprana (SAT), es permitir alertar y dirigir en caso de desastres naturales o climatológicos, para que comunidades amenazadas puedan actuar a tiempo y de manera apropiada para reducir la posibilidad de pérdidas de vidas humanas, daños a la propiedad y al ambiente y pérdida del sustento [20].

Existen estudios realizados [21], [34], en donde se quiere minimizar mediante la prevención de riesgo los daños que sufre la sociedad al tener que enfrentarse a desastres naturales como son: los desprendimientos de tierra, terremotos, inundaciones, etcétera. Estos fenómenos se originan comúnmente en ríos de montaña en donde existen fuertes vertientes, zonas con actividad volcánica.

### 2.2 Sistemas de análisis de sentimientos

Los sistemas de análisis de sentimientos o también conocidos como “minería de opinión”, utilizan el procesamiento del lenguaje natural y la similitud semántica, para proporcionar información del estado de ánimo o grado de satisfacción de un determinado usuario analizando las frases, comentarios, hashtags, etcétera, que pública.

La minería de opiniones es una reciente sub-disciplina entre la recuperación de información y la lingüística computacional. Esta sub-disciplina no trata de detectar el tema abordado en un texto sino las opiniones y sentimientos expresados en el mismo [35]. Determina si los comentarios en foros online, blogs o redes sociales sobre un tema en particular (producto, libro, película, etc.) son positivos, negativos o neutros. Las opiniones son muy importantes cuando alguien desea conocer la opinión de los otros antes de tomar una decisión.

Tradicionalmente se pueden distinguir dos metodologías principales en minería de opiniones. Por un lado hay enfoques para minería de opiniones que tienen como objetivo clasificar documentos enteros con polaridad positiva o negativa.

Cuando se habla de minería de opiniones, también llamada análisis del sentimiento, se refiere a la aplicación de una serie de técnicas del procesamiento del lenguaje natural, lingüística computacional y minería de textos, que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios, como comentarios en blogs u opiniones en revistas de productos [36], [37], [38]. Con este tipo de tecnologías es posible extraer un valor tangible y directo a partir de un comentario en lenguaje natural. Se podrían monitorizar y gestionar las opiniones sobre personas o entidades en foros y redes sociales concretas. Así como abordar temas como las tendencias [39].

Existen investigaciones realizadas[40], [41], [42], que para solventar el problema de la gran cantidad de información generada en los medios sociales, hace uso de técnicas de procesamiento del lenguaje natural (PLN). Estos trabajos estudian y proponen métodos para determinar la polaridad sobre opiniones, hashtags, tweets, así como determinar el tópico del mismo. Además de las investigaciones mencionadas, también hay herramientas que son capaces de analizar un tweet y determinar si son de carácter positivo o negativo, tales como:

* Mr. Tuit (<http://www.mrtuit.com/>).

### 2.3 Propuesta de proyecto

#### 2.3.1 Red social del proyecto.

En nuestro proyecto tras analizar las distintas redes sociales, nos hemos decantado por trabajar con la red social Twitter. Al tener miles de usuarios compartiendo información diariamente a través de los tweets, hashtags, entre otros, se manejan grandes volúmenes de datos con diversidad de ámbitos como: la política, el turismo, opiniones digitales, etcétera. Para abarcar el procesamiento de todos los datos generados en la redo social Twitter utilizaremos el lenguaje R, ya que, es un lenguaje muy potente a la hora de tratar con la minería de datos. También trabajaremos con procesamiento del lenguaje natural (PLN), para poder analizar los datos extraídos de la red social. Utilizando mecanismos como los sistemas de alerta y sistemas de análisis de sentimientos para el tratamiento de los datos.

#### 2.3.2 Nuestra Propuesta

Para abarcar el procesamiento de todos los datos generados en las redes sociales por los usuarios, nace la necesidad de utilizar mecanismos como los sistemas de alerta y sistemas de análisis de sentimientos. Los sistemas de alerta, son sistemas que ante uno o varios eventos, son capaces de generar una notificación o alerta, para indicar a un usuario o grupo de usuarios sobre un suceso, acontecimiento. Los sistemas de análisis de sentimientos o también conocidos como “minería de opinión”, utilizan el procesamiento del lenguaje natural y la similitud semántica, para proporcionar información del estado de ánimo o grado de satisfacción de un determinado usuario. Estos sistemas, aplicados en el sector turístico, proporcionarían un medio rápido y efectivo a la hora de comunicar a las personas, por ejemplo, acerca de la situación climatológica generando una alerta por lluvia en el caso de los sistemas de alerta. Y en el caso de los sistemas de análisis de sentimientos, opiniones positivas o negativas dependiendo del léxico empleado, basándose en las propias experiencias de los usuarios.

Nuestra propuesta consiste realizar una integración de un sistema de alerta y un sistema de análisis de sentimientos orientando en el ámbito del turismo. Para ello trabajaremos con un grupo de tecnologías descritas en el Capítulo 1, sección 1.5. Para la obtención y la extracción de los datos utilizaremos como fuente de información a la red social Twitter.

En esta propuesta, se quiere elaborar un sistema de alerta y un sistema de análisis de sentimientos, aplicados en el sector turístico, para así proporcionar un medio rápido y efectivo a la hora de comunicar o alertar a las personas de algún suceso. Por ejemplo, acerca de la situación climatológica generando una alerta por lluvia en el caso de los sistemas de alerta. Y en el caso de los sistemas de análisis de sentimientos, opiniones positivas o negativas dependiendo del léxico empleado, basándose en las propias experiencias de los usuarios.

## Capítulo 3

## Análisis de requisitos y caso de uso.

En el capítulo anterior, definíamos en qué consistían los sistemas de alerta y de análisis de sentimiento, así como la propuesta de sistema para el proyecto.

En este capítulo se muestran los requisitos funcionales y no funcionales establecidos para la integración de sistemas, desarrollado en este Trabajo de Fin de Grado. También se recogen diagramas de casos de uso de la aplicación y las fuentes utilizadas.

### 3.1 Requisitos

#### 3.1.1 Servidor

* Requisitos funcionales:
  + Recepción de peticiones de cliente: el servidor será capaz de alojar la aplicación.
  + Publicación de los datos: el servidor tendrá acceso a una lista en donde se verán reflejados la información buscada.
* Requisitos no funcionales:
  + Acceso a fichero con datos: el servidor ha de acceder a los datos (usuarios, tweets) realizados en la búsquedas para utilizarlo en el cumplimiento de los requisitos funcionales.

#### 3.1.2 Cliente

* Requisitos funcionales:
  + Importación de conjuntos de datos del servidor: la aplicación cliente deberá ser capaz de acceder a la lista de datos que proporciona el servidor bajo los ficheros.csv.
  + Gráfica de tweets: la aplicación deberá mostrar una gráfica distinguiendo en tres colores, las distintas clasificaciones de los tweets, según su naturaleza.
  + Tabla de información: la aplicación deberá proporcionar una tabla, en donde se reflejan los datos (usuario, tweets) sobre los que se realiza el análisis.
  + Formulario para generación de búsquedas: deberá existir un formulario que el usuario deberá rellenar con los siguientes campos:
    - Búsqueda: con el término que se desea buscar.
    - Número de tweets: entero que establecerá el número de tweets sobre el que se realizara la búsqueda.
    - Tipo de alerta: en donde el usuario elegirá el tipo de notificación que desea recibir.
* Requisitos no funcionales
  + Plataformas: será una aplicación web, compatible con los navegadores más populares.
  + Interfaz: la interfaz deberá ser accesible y usable.

### 3.2 Diagramas de casos de uso

* Servidor

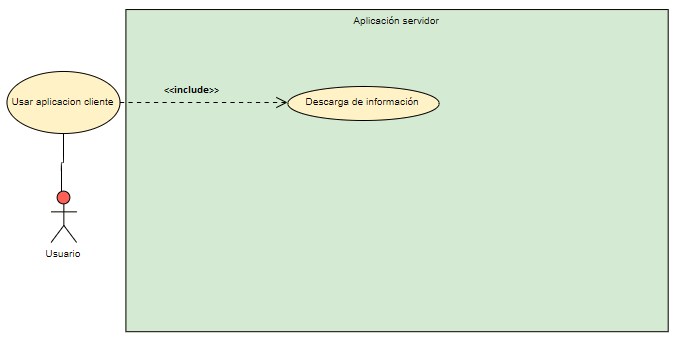


Figura 4: Caso de uso del servidor.

* Cliente

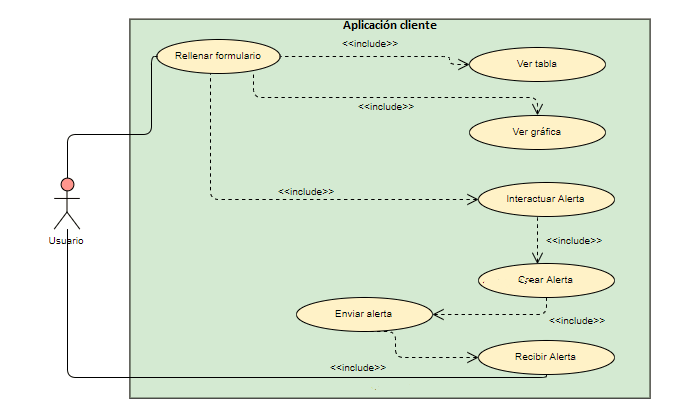


Figura 5: Caso de uso del cliente

## Capítulo 4

## Integración de los sistemas de alerta y de análisis de sentimientos propuestos.

En el capítulo anterior, hicimos un análisis de los requisitos para la integración de sistemas, así como la aplicación de un caso de uso.

En este capítulo, explicaremos detalladamente cómo hacer una integración de estos dos sistemas. Empezando en como seleccionamos y extraemos los datos, para su posterior procesamiento y formato de salida.

### 4.1 Selección de datos.

La fuente de donde obtendremos los datos es la red social Twitter, ya que, abarca mucha información de miles de usuarios. Esta información está representada en distintos formatos, texto, imágenes, videos, etc. Nosotros utilizaremos la información en formato texto, porque a la hora de procesar estos datos, al estar en formato plano y digitalizado, el procesamiento del lenguaje natural, será más provechoso. Por este motivo, el contenido que recogeremos de la red social, serán: hashtags, tweets, retweets y menciones.

### 4.2 Extracción de datos.

Para obtener los datos de Twitter lo haremos, mediante el lenguaje de programación R bajo el entorno de trabajo RStudio. Para extraer estos datos, se deben seguir unos pasos, que se detallarán a continuación.

Crear acceso a Twitter.

1. Cuenta en Twitter.

Es necesario disponer de una cuenta en Twitter, la que utilizaremos para el desarrollo de la aplicación es: Sistemas de Alerta y Análisis de Sentimientos “SAYAS” (<https://twitter.com/SAYAS91995364>).

1. Registro de una API en Twitter.

Para crear la aplicación, debemos registrarla en (<https://dev.twitter.com/apps>), pulsar “crear new app” y cumplimentar los datos de registro. El nombre de la aplicación que hemos registrado es: SAYAS-2017.

Autenticación de Twitter con R.

Para la autenticación en Twitter, haremos uso del código que se detalla a continuación:

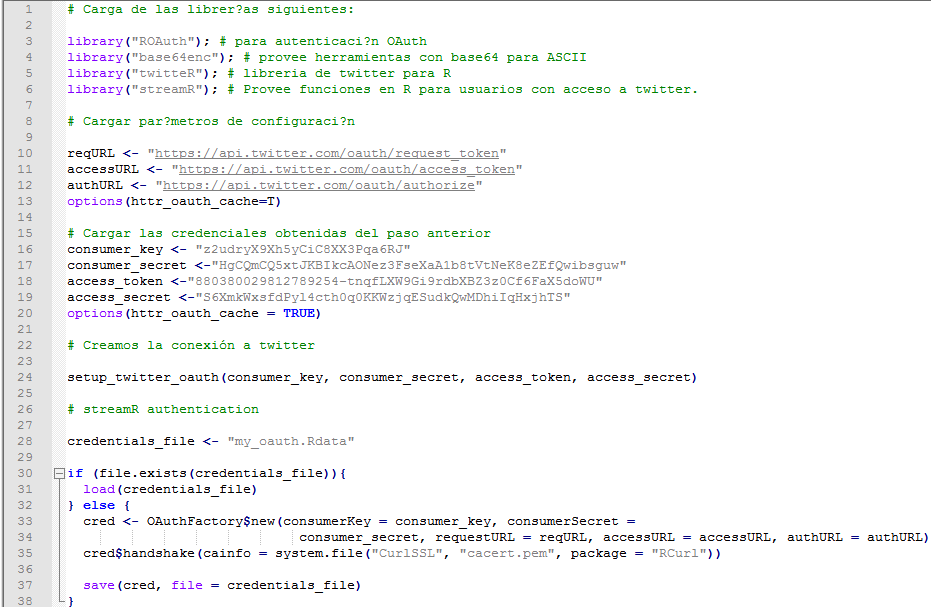


Figura 6: Código para autenticación con OAuth.

Una vez ejecutado el código, se abrirá una pestaña en el navegador, solicitando el acceso a la aplicación, pulsaremos “permitir acceso” y nos devolverá un PIN. Este PIN, es el que introducimos por consola en Rstudio, y con eso se nos creará el fichero “My\_Oauth.data” para el login.

Construccion del dataframe

Para la construcción del dataframe, hemos utilizado la función searchTwitter(valor, tamaño, lenguaje, localizacion), la cual realiza una búsqueda según el valor del paramétro “input$busqueda” que se le pasa en el formulario de la aplicación.En cuanto al tamaño máximo de retorno, lo hemos establecido en 100, el idioma en español “es” y como se trata de una aplicación a nivel de Canarias, hemos pasado su localización en coordenadas geográficas. Posteriormente con la función twLisToDF hemos transformado la lista devuelta en la búsqueda en un dataframe y seleccionando, las variables “text” y “screenName”, hemos dado forma al dataframe, el cual esta compuesto por una lista de usuarios y sus correspondients tweets.

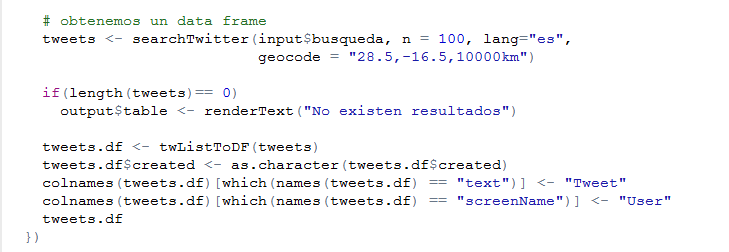


Figura 7: Dataframe de extracción de datos

### 4.3 Tratamiento de los datos obtenidos

Como los datos obtenidos en la fase extracción son de diferentes tipos, recibirán un procesamiento para que así, la información pueda manipularse de manera más sencilla. Para ello, realizaremos una serie de pasos: 1. Volcar datos obtenidos del dataframe en un fichero CSV, 2. Limpiar los datos obtenidos en la extracción.

1. Volcar datos obtenidos del dataframe en un fichero CSV.

Una vez tenemos nuestro dataframe, metemos los datos en un fichero CSV para la posterior depuración de los datos.

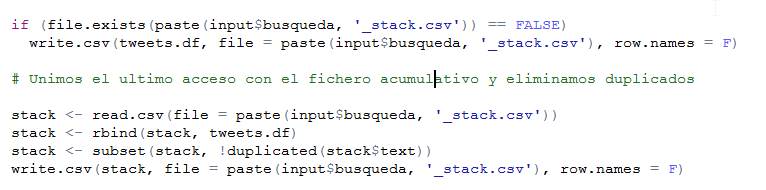


Figura 8: Datos en fichero CSV

1. Limpiar los datos obtenidos en la extracción,

Los datos almacenados en el fichero CSV, contienen elementos que no nos interesan por lo que es conveniente eliminarlos. Como por ejemplo links, símbolos de puntuación, números y duplicados. Para eliminar los duplicados haremos uso de la función subset(), combinado con la función ¡duplicated().También utilizaremos el método gsub (“expresión”,”replace”,variable). Esta función, permite eliminar contenido de dentro de una variable, le pasamos tres parámetros. El primero es la expresión a buscar y el segundo elemento por lo que reemplazar y tercero, la variable en donde se hará esa búsqueda. Para obtener unos datos “limpios”, hemos eliminado los links, los signos de puntuación, números, transformado las palabras a minúsculas y separado las palabras por espacios.

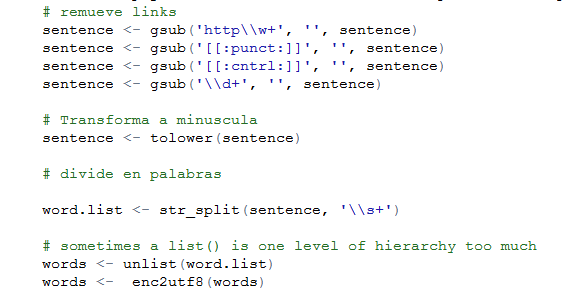


Figura 9: Limpieza de datos.

### Análisis de los datos

Una vez obtenidos los datos de la red social Twitter y posteriormente de realizar un tratamiento a estos, en donde se realizan distintos procedimientos para así obtener un texto que se le pueda aplicar el siguiente análisis.

Para realizar en análisis de sentimientos de los tweets obtenidos en la fase de extracción de datos, la estrategia empleada, consiste en contrastar cada una de las palabras de cada mensaje con las contenidas en los ficheros positive-words.txt y negative-words.txt. Estos archivos, funcionan como una “Base de Datos” los cuales encierran palabras en el idioma “Español” con significado positivo y negativo, respectivamente.

El objetivo que se persigue con esto, es el de establecer una calificación a cada mensaje “tweet” que pueda ser representada. Así pues, las puntuaciones positivas equivalen a tweets cuyos mensajes tienen connotación positiva (mayor a cero), las negativas (menor que cero) tienen connotación negativa y las notas que sean “0” son tweets neutros, los cuales no entran dentro de los dos grupos anteriormente mencionados.

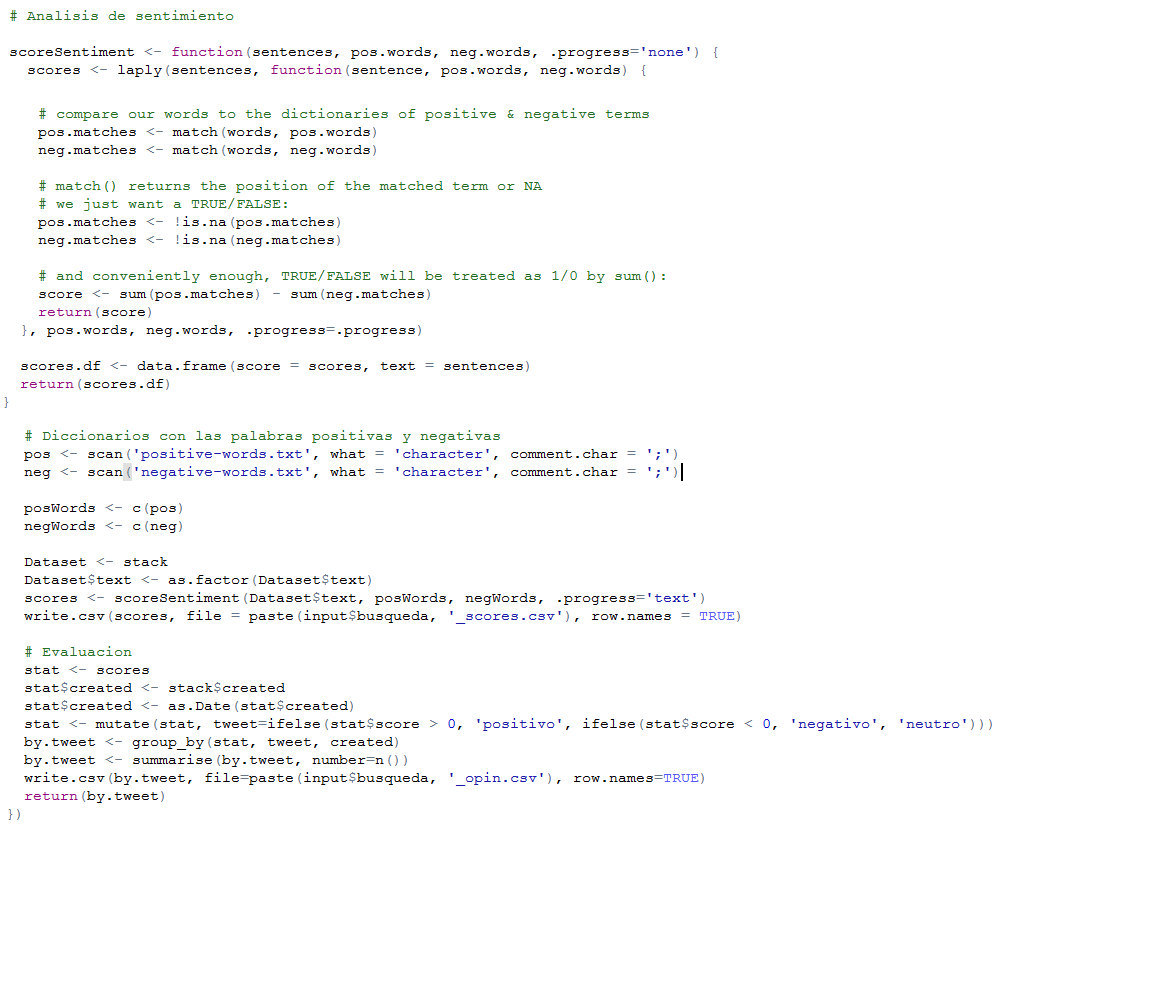


Figura 10: Análisis de sentimientos.

### Alertas enviadas

A la hora de realizar las alertas se ha optado por realizarlas vía correo electrónico. Mediante la utilización de la librería “mailR”, la cual permite el envío de correos electrónicos con lenguaje R.

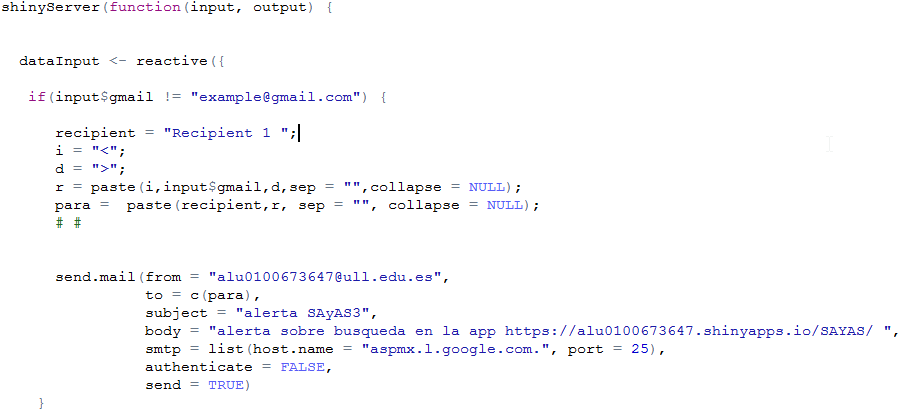


Figura11: Código para envío de alertas en el servidor.

### Librerías utilizadas para el desarrollo de la app

A continuación se detalla una tabla en donde hemos clasificado las librerías utilizadas, según el tipo de uso que se le ha dado.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Conexión y extracción de datos | Tratamiento de datos | | Sistema de Alerta | Sistema de Análisis de Sentimientos | Visualización y despliegue de la app |
| ROAuth | | dplyr | mailR | leaflet | shiny |
| base64enc | | tm |  | stringr | shinythemes |
| twitteR | | stringr |  |  | base64enc |
| streamR | | reader |  |  | ggthemes |
| ROAuth | | plyr |  |  | rsconnect |
|  | |  |  |  | ggplot |
|  | |  |  |  | RPushbullet |

Figura 12: Librerías utilizadas

### 4.7 La aplicación con ShinyApp

Para la implementación de la aplicación web, se ha utilizado la librería “shiny”, que facilita la creación de aplicaciones web interactivas directamente desde R. También tiene la posibilidad de extender sus aplicaciones con temas CSS, HTML widgets y acciones JavaScript. Para el despliegue, se ha utilizado el servicio de hosting que ofrece ShinyApps.

* Código para la visualización de la aplicación

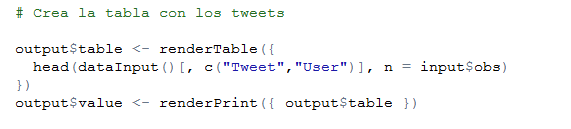


Figura 13: Tabla de análisis en el servidor

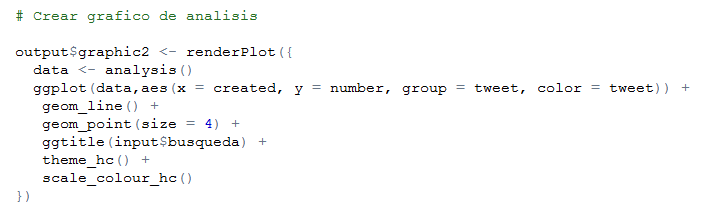


Figura 14: Gráfico de análisis en el servidor

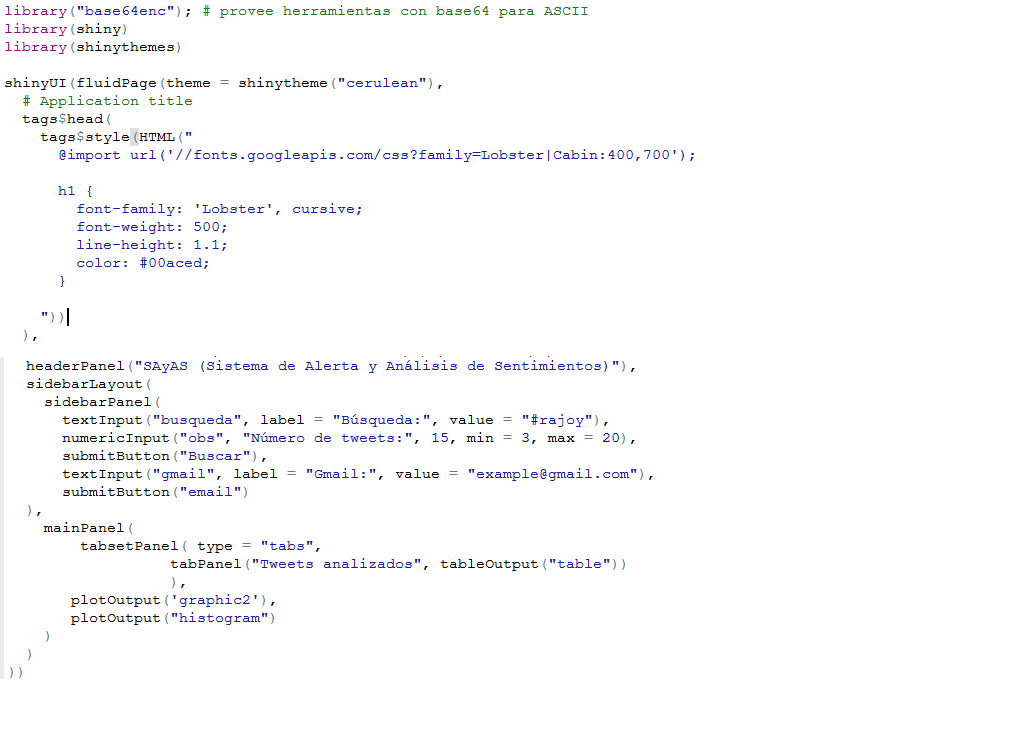


Figura 15: Estructura de la aplicación

Apariencia de la aplicación

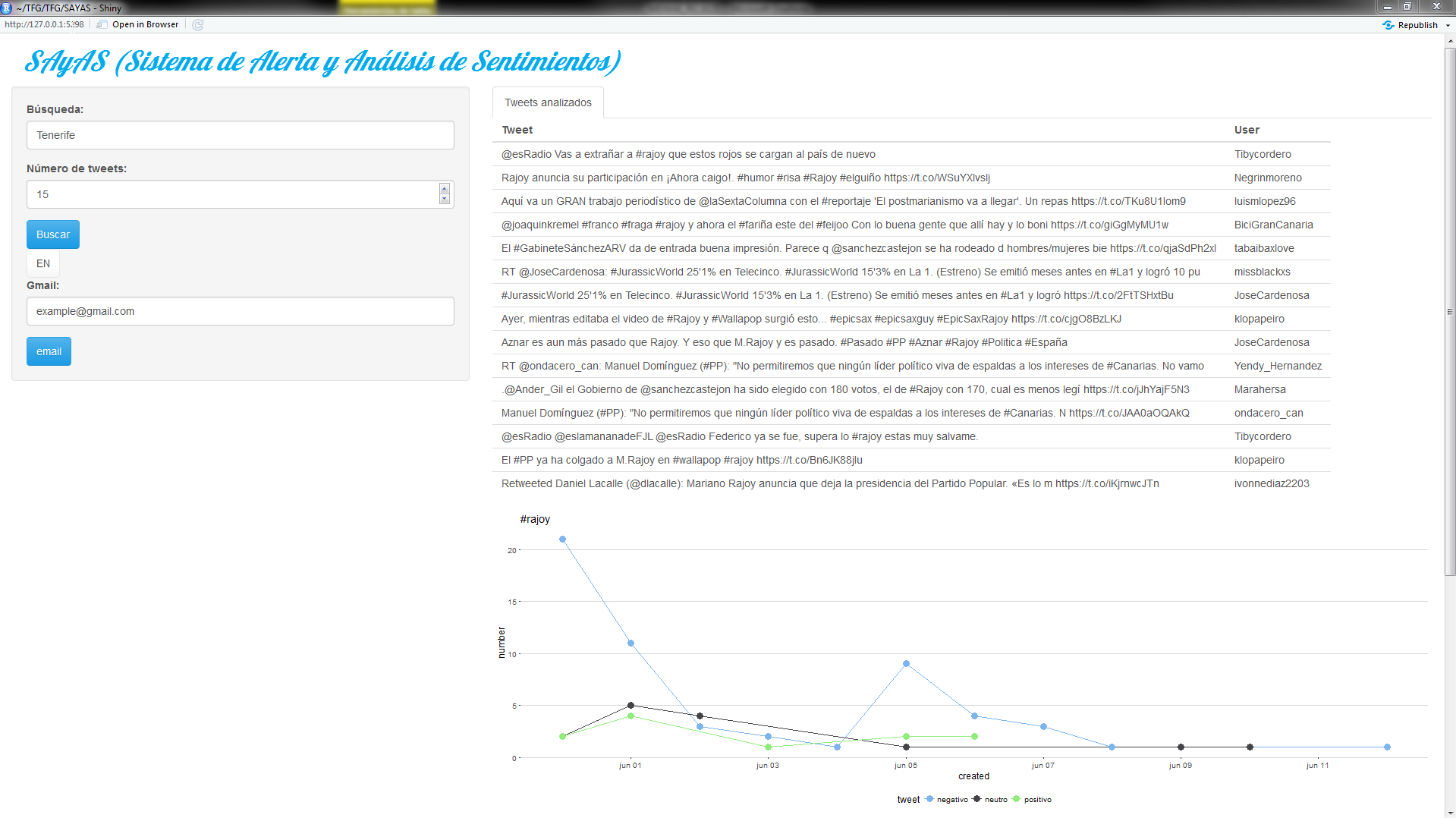


Figura 16: La aplicación

## Capítulo 5

## Conclusiones y líneas futuras.

En este capítulo se describen las conclusiones obtenidas posteriormente a la realización del Trabajo Fin de Grado, con el propósito de revisar los objetivos que fueron establecidos en las distintas secciones durante su realización y para comprobar el grado de satisfacción en cada uno de ellos. A posteriori de este análisis, se cotejan algunas líneas futuras para otros posibles Trabajos de Fin de Grado, en los que se pudiera optimizar o ampliar las funcionalidades del sistema de alerta y de análisis de sentimientos como resultado del trabajo presente.

Este sistema se ha basado en una integración de sistemas. En concreto, en un sistema de alerta y un sistema de análisis de sentimientos. Este sistema, denominado “SayAS (Sistema de Alerta y Análisis de Sentimientos), permite a los usuarios, a través de una aplicación web, obtener un análisis a cerca de una palabra buscada, así como recibir una notificación de esa búsqueda.

### 5.1 Conclusiones

En el capítulo 1 de la presente memoria, se han expuesto una serie de objetivos a perseguir, cuyos resultados y conclusiones son los siguientes:

1. Se ha logrado una recopilación lo suficientemente completa y concisa de la evolución de lo que se conoce como Internet, como surgieron las redes sociales, redes colaborativas y su similitud con las redes sociales.
2. Se han analizado diferentes metodologías para los sistemas de alerta, así como para los sistemas de análisis de sentimientos.
3. Se ha diseñado e implementando un sistema de alerta y un sistema de análisis de sentimientos con los datos obtenidos de la red social Twitter.
4. Se ha llevado a cabo la integración en una sola herramienta de un sistema de alerta y un sistema de análisis de sentimientos. Así como, su correcto funcionamiento.
5. Se ha propuesto distintos casos de uso en donde se puede emplear la herramienta desarrollada.
6. Se ha documentado el código desarrollado y creado la memoria final de este Trabajo de Fin de Grado.

Por lo tanto, como se han cumplido todos los objetivos propuestos, se puede concluir con que este Trabajo de Fin de Grado se ha llevado a cabo con éxito.

Durante su realización, fue imprescindible estudiar y comprender como funcionan cada uno de los sistemas integrados, y en que consiste un sistema de alerta, y un sistema de análisis de sentimientos. También el procesamiento del lenguaje natural (PLN), que tiene un amplio alcance en diferentes aplicaciones, su arquitectura a diferentes niveles y los componentes necesarios para usar y manejar de manera más provechosa la información plasmada en los textos. Por otro lado, fue importante aprender a usar las herramientas como RStudio, lenguaje R, y comprender muchas de sus librerías como Shiny, twitterR, entre otras.

En este Trabajo de Fin de Grado, se pudo apreciar la potencia que tienen los sistemas de alerta, para generar las notificaciones o alertas en situaciones extremas, o simplemente para un aviso de algún acontecimiento que este ocurriendo en ese preciso instante. En cuanto a los de sistemas de análisis de sentimientos o también conocidos como minería de opinión, son sistemas muy complejos, ya que, a la hora de procesar y tratar la información son capaces obtener un resultado favorable o no dependiendo de la información contrastada.

Estamos en la era de la información y de los grandes volúmenes de datos, todo gira en torno de las redes sociales y la opinión digitaliza. Este recurso es de vital importancia, no solo para obtener información acerca de los usuarios en las redes, sino para conocer a las personas y poder así, procesar dicha información y transformarla en un bien que pueda mover un sector como es el sector turístico en Canarias, y paralelamente al sector económico. Construir un sistema de tales características ha sido muy enriquecedor e interesante para mi, la experiencia y conocimientos adquiridos han sido clave, para iniciarme en este mundo del Big Data.

### 5.2 Líneas Futuras

Como líneas futuras, se detalla una lista de líneas futuras de trabajo para la integración de sistemas, resultante de este Trabajo de Fin de Grado. Estas consisten en nuevas características y funcionalidades que podrían estudiarse para mejorar y optimizar así el trabajo realizado en este proyecto:

1. Otros formatos de alerta: se podría añadir una funcionalidad en donde el usuario además de recibir una alerta vía Gmail, pudiera recibir una alerta vía Twitter o SMS.
   1. Envío de alerta similar a la búsqueda cada equis tiempo. Junto a una Gráfica con el análisis.
2. Sistema de geolocalización: además de los sistemas ya integrados, se podía incluir otro sistema de geolocalización, en donde dependiendo de la ubicación del usuario, pues reciba una alerta del tiempo que pueda hacer en esa zona. Otra funcionalidad de un sistema de geolocalización es la posibilidad de añadir una dirección en donde se quiera ir y que este le indique la experiencia de otros usuarios y el clima en la zona.
3. Añadir otros idiomas: como es una aplicación orientada al ámbito de turismo, y como gran parte de los turistas en Canarias son extranjeros, sería una buena idea, añadir la opción de traducir la interfaz y los resultados de la aplicación a otros idiomas como el Inglés, Alemán o Portugués.

## Capítulo 6

Summary and Conclusions

## Capítulo 7

## Presupuesto

En este capítulo se realiza una estimación aproximada del proyecto, en función de las horas dedicadas y de los recursos utilizados para su elaboración. El presupuesto estimado para la realización de este proyecto se basa en la duración de horas de este, junto a su implementación. Cada hora es estimada en un total de 10,5 euros.

#### 7.1 Presupuesto del trabajo realizado

Tiempo empleado

|  |  |  |  |
| --- | --- | --- | --- |
| Objetivos | Descripción | Horas totales | Coste |
| Estudio del estado del arte | Estudio del estado del arte relacionado con las redes sociales, así como los distintos sistemas de alerta y de análisis de sentimientos en la red social twitteR. | 120 | 1.260 |
| Análisis de metodologías | Análisis de métodos para sistemas de alerta y sistemas de análisis de sentimientos utilizando a la red social Twitter como fuente de información. | 80 | 840 |
| Diseño e implementación | Diseñar e implementar un sistema de alerta y un sistema de análisis sentimientos con los datos propuestos | 150 | 1575 |
| Testeo, validación e integración | Testear, validar e integrar en una sola herramienta, un sistema de alerta y un sistema de análisis de sentimientos. | 70 | 735 |
| Caso de uso | Proponer un caso de uso para la herramienta que tendrá integrada un sistema de alerta y un sistema de análisis de sentimientos | 50 | 525 |
| Documentación | Elaboración de documentación técnica, entregables y memoria del proyecto. | 140 | 1.470 |
| Total | | 610 | 6405 |

Tabla 7.1. Tabla resumen de los Tipos.

**Hardware utilizado**

|  |  |  |
| --- | --- | --- |
| **Dispositivo** | **Descripción** | **Coste(€)** |
| ASUS A53S | Ordenador portátil | 425 |

Tabla 7.2 Presupuesto hardware utilizado

El coste estimado del ordenador portátil, no se suma al presupuesto de tiempo invertido, ya que, ya se disponía de este antes de la realización del proyecto.

**Software utilizado**

|  |  |
| --- | --- |
| Software | Coste(€) |
| GitHub | 0 |
| Windows 10 | 135 |
| Google Chrome, Mozilla Firefox, Microsoft Edge | 0 |
| Servidor ShinyApps | 0 |

Tabla 7.3: Presupuesto del software utilizado

El sistema operativo Windows 10, está incluido en el ordenador utilizado para la realización del Trabajo de Fin de Grado, por lo que su coste no afecta al presupuesto real de tiempo invertido.

**Presupuesto total del proyecto**

|  |  |
| --- | --- |
| **Recursos** | **Coste (€)** |
| Tiempo empleado | 6.406 |
| Software | 0 |
| Hardware | 0 |
| Total | 6406 |

Tabla 7.4. Tabla resumen del presupuesto.

## Referencias

|  |  |
| --- | --- |
| [1] | E. B. Britannica Academic, «Internet,» 2016. |
| [2] | T. Berners-Lee y R. Cailliau, «WorldWideWeb: Proposal for a HyperText project,» *Retrieved on February,* vol. 26, p. 2008, 1990. |
| [3] | K. Kapp y T. O’Driscoll, «Learning in 3D,» *Adding a new dimension to enterprise learning and collaboration, Pfeiffer, San Francisco, CA,* 2010. |
| [4] | G. Domínguez Fernández, «LA EDUCACIÓN SOCIAL Y LA WEB 2.0: NUEVOS ESPACIOS DE INNOVACIÓN E,» *Pixel-Bit. Revista de Medios y Educación,* pp. 105-114, 2009. |
| [5] | M. d. C. {Llorente Cejudo, «Assessing personal learning environments,» *University of Alicante,* 2013. |
| [6] | T. O’REILLY, What is web 2.0., 2005. |
| [7] | X. Ribes, «La Web 2.0. El valor de los metadatos y de la inteligencia colectiva,» *Telos,* vol. 73, pp. 36-43, 2007. |
| [8] | J. J. Flores Cueto, J. J. Morán Corzo y J. J. Rodríguez Vila, «Las redes sociales,» *Universidad de San Martín de Porres,* pp. 1-15, 2009. |
| [9] | C. Cobo y H. P. Kuklinski, Planeta Web 2.0.: Inteligencia colectiva o medios fast food, LMI, 2000. |
| [10] | L. B. Prato, Aplicaciones Web 2.0: Redes Sociales, Buenos Aires: Eduvim-Editorial Universitaria Villa María, 2010. |
| [11] | J. Cabero, J. Barroso, M. Llorente y V. Marín, *Las redes sociales como instrumentos para la formación. Percepciones de los alumnos universitarios hacia el trabajo en grupo,* Sevilla, España: Grupo de Investigación Didáctica. Universidad de Sevilla, 2013. |
| [12] | Charles, N. B. ELLISON y C. LAMPE, «Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. Journal of Applied Developmental Psychology,» 2008. |
| [13] | J. J. De Haro Ollé, «Redes sociales para la educación,» *Anaya Multimedia. Madrid,* 2010. |
| [14] | J. Cabero, «Las redes sociales en el entramado educativo de la Web. 2.0,» *E. Navas, Web,* vol. 2, pp. 11-28, 2012. |
| [15] | L. Abramson, «Can social networking keep students in school,» *NPR: Morning edition,* 2011. |
| [16] | M. S. Granovetter, «The strength of weak ties,» *American journal of sociology,* vol. 78, nº 6, pp. 1360-1380, 1973. |
| [17] | I. N. De Ledo, «Las redes sociales,» *Revista Venezolana de Oncolog{\'\i}a,* vol. 23, nº 3, p. 133, 2011. |
| [18] | «Observatorio Tecnológico,» 2012. [En línea]. Available: http://recursostic.educacion.es/observatorio/web/en/internet/web-20/1043-redes-sociales?start=3. [Último acceso: 26 04 2017]. |
| [19] | L. M. Camarinha-Matos y H. Afsarmanesh, «Collaborative networks: a new scientific discipline,» *Journal of intelligent manufacturing,* vol. 16, nº 4-5, pp. 439-452, 2005. |
| [20] | Y. C. Betancourt y R. C. Menéndez, «Sistemas de Alerta Temprana Escolar en caso de inundaciones en la parroquia Panaquire, estado Miranda. Una propuesta educativa ambiental en gesti{\'o}n de riesgo,» *REVISTAS DE INVESTIGACIÓN,* vol. 38, nº 82, 2014. |
| [21] | J. Rosales, F. Campos, V. Córdova, C. Matos, J. Corimanya, W. Palomino y J. Castillo, «UNA PROPUESTA DE MODELO DE DIFUSIÓN DE ALERTA TEMPRANA PARA DESLIZAMIENTOS DE TIERRA EN EL PERÚ USANDO LA TELEVISIÓN DIGITAL TERRESTRE,» *Revista Cientifica TECNIA,* vol. 25, nº 1, p. 53, 2017. |
| [22] | H. ALDRICH y D. A. WHETTEN, «Organization-sets, action-sets, and networks: Making the most of simplicity. Handbook of organizational design,,» pp. 385-408, 1981. |
| [23] | S. P. Borgatti y P. C. Foster, «The Network Paradigm in Organizational Research: A Review and Typology,» *Journal of Management,* vol. 29, nº 6, pp. 991-1013, 2003. |
| [24] | R. D. Franco Pereyra, «Arquitectura, metodología y plataforma tecnológica para la ingeniería y operación de redes colaborativas. Una aproximación basada en servicios digitales,» pp. 41-69, 2012. |
| [25] | A. Cortez Vásquez, J. Pariona Quispe y A. Huayna, «Procesamiento de Lenguaje Natural,» *Revista de investigación de Sistemas e Informática,* vol. 6, nº 2, pp. 45-54, 2014. |
| [26] | A. Gelbukh, «Procesamiento de lenguaje natural y sus aplicaciones,» *Korpus Sapiens. Sociedad Mexicana de inteligencia artificial,* vol. 1, 2010. |
| [27] | Rstudio, *Rstudio.* |
| [28] | «Rproject,» [En línea]. Available: https://www.r-project.org/about.html. [Último acceso: 2017]. |
| [29] | A. G. y Silvia González, «R Development Core Team,» The address of the publisher, 2000. |
| [30] | J. Mulero, «"Aplicaciones interactivas diseñadas con Shiny.",» 2016. |
| [31] | F. J. Lopez-Pellicer, «GitHub como herramienta docente.Actas de las XXI Jornadas de la Enseñanza Universitaria de la Informática,» pp. 66-73, 2015. |
| [32] | OVTT, *OVTT.* |
| [33] | L. C. Bolaños Martínez y F. J. C. Caicedo, «Evaluación de la implementación de sistemas automatizados de alerta temprana en tiempo real dirigidos a disminuir los riesgos que representan las amenazas por inundaciones y avenidas torrenciales en la subcuenca Rio Molino, municipio de Popayán, departamento del Cauca,» 2016. |
| [34] | M. A. A. Coll, «Sistemas de Alerta Temprana (SAT) para la Reducci{\'o}n del Riesgo de Inundaciones S{\'u}bitas y Fen{\'o}menos Atmosf{\'e}ricos en el {\'A}rea Metropolitana de Barranquilla,» *Scientia et technica,* vol. 18, nº 2, pp. 303-308, 2013. |
| [35] | A. Esuli y F. Sebastiani, «Determining the Semantic Orientation of Terms Through Gloss Classification,» de *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, New York, NY, USA, 2005. |
| [36] | F. L. Cruz, J. A. Troyano, F. Enrí­quez, F. J. Ortega y C. G. Vallejo, «Long autonomy or long delay?’The importance of domain in opinion mining,» *Expert Systems with Applications,* vol. 40, nº 8, pp. 3174-3184, 2013. |
| [37] | A. Moreo, M. Romero, J. Castro y J. M. Zurita, «Lexicon-based comments-oriented news sentiment analyzer system,» *Expert Systems with Applications,* vol. 39, nº 10, pp. 9166-9180, 2012. |
| [38] | M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez y L. Ureña-López, «Experiments with SVM to classify opinions in different domains,» *Expert Systems with Applications,* vol. 38, nº 12, pp. 14799-14804, 2011. |
| [39] | I. Peñalver Martínez, «Minería de opiniones basada en características guiada por ontología,» *Proyecto de investigación:,* 2015. |
| [40] | M. d. P. Salas-Zárate, «Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computacional and Mathematical Methods in Medicine,» 2017. |
| [41] | F. Pla y L.-F. Hurtado, «ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter,» de *XXIX Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS*, 2013. |
| [42] | J. M. ;. C. M. F. ;. O. F. J. ;. T. J. J. A. Cotelo, «Explorando Twitter mediante la Integración de Información Estructurada y No Estructurada.Procesamiento del lenguaje natural,» 2015. |
| [43] | A. C. Vásquez, J. P. Quispe, A. M. Huayna y others, «Procesamiento de lenguaje natural,» *Revista de investigaci{\'o}n de Sistemas e Inform{\'a}tica,* vol. 6, nº 2, pp. 45-54, 2009. |
| [44] | J. Musser y T. O’reilly, «Web 2.0,» *Principles and Best Practices.[Excerpt]. oO: O'Reilly Media,* 2006. |
| [45] | J. M.-M. K. L.-O. H. L.-A. M. Á. R.-G. María del Pilar Salas-Zárate y R. Valencia-García, «Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach, Computational and Mathematical Methods in Medicine,» 2017. |
| [46] | A. Gelbukh, «Procesamiento de Lenguaje Natural y sus Aplicaciones,» *Komputer Sapiens,* 2010. |
| [47] | N. B. Ellison y others, «Social network sites: Definition, history, and scholarship,» *Journal of Computer-Mediated Communication,* vol. 13, nº 1, pp. 210-230, 2007. |
| [48] | D. J. Brass, J. Galaskiewicz, H. R. Greve y W. Tsai, «Taking stock of networks and organizations: A multilevel perspective,» *Academy of management journal,* vol. 47, nº 6, pp. 795-817, 2004. |
| [49] | J. M. d. N. S. M. L. Silva, «Gestão de informação de desempenho em redes colaborativas de organizaçoes,» 2012. |
| [50] | J. C. Almenara, J. B. Osuna, M. d. C. L. Cejudo y C. Y. Cabrera, «Redes sociales y Tecnologías de la Información y la Comunicación en Educación: aprendizaje colaborativo, diferencias de g{\'e}nero, edad y preferencias,» *Revista de Educación a Distancia,* nº 51, 2016. |
| [51] | M. d. C. Llorente Cejudo, «Assessing personal learning environments (PLEs). An expert evaluation,» 2013. |
| [52] | C. C. Romaní y H. P. Kuklinski, «PLANETA WEB 2.0,» *Inteligencia colectiva o medios fast food. Group de Recerca d’Interaccions Digitals, Universitat de Vic,* 2007. |